Marginal and Conditional Distribution Estimation from Double-Sampled Semi-Competing Risks Data

Menggang Yu[1,*] and Constantin T. Yiannoutsos[2]

[1]Department of Biostatistics & Medical Informatics
K6/446 CSC 600 Highland Ave.
Madison, WI 53792-4675, USA.

[2]Department of Biostatistics,
Indiana University School of Medicine,
410 West 10th Street, Suite 3000,
Indianapolis, IN 46202, USA.

* Correspondence to: meyu@biostat.wisc.edu

May 19, 2012

Abstract

Subject dropout is a vexing problem for any biomedical study, especially when the dropout subjects differ from the non-dropout subjects in terms of the main outcome(s). Usual statistical method that intends to correct estimation bias related to this phenomenon involves unverifiable assumptions about the dropout mechanism. We consider a unique cohort study in Africa that uses an outreach program to ascertain the mortality information vital status for dropout subjects. These data can be used to identify a number of relevant distributions. However only a subset of dropout subjects were followed, vital status ascertainment was incomplete. We use semi-competing risk methods as our analysis framework to address this specific case where the terminal event is incompletely ascertained and consider various procedures for estimating the marginal distribution of dropout and the marginal and conditional distributions of survival. We also consider model selection and estimation efficiency in our setting. Performance of the proposed methods is demonstrated via simulations, asymptotic study, and analysis of the study data.

KEY WORDS: Dropout, double sampling, Copula model, semi-competing risks.

This is the author's manuscript of the article published in final edited form as:
Yu, M., & Yiannoutsos, C. T. (2015). Marginal and Conditional Distribution Estimation from Double-sampled Semi-competing Risks Data. Scandinavian Journal of Statistics, 42(1), 87-103. http://dx.doi.org/10.1111/sjos.12096

1 Introduction

Competing risk data are common for time-to-event outcomes (Klein and Moeschberger 1997). In particular, the usual censored failure time data can be viewed as competing risk data when censoring is regarded as a competing cause of failure. More generally, study subjects may experience a number of distinct failure types. It is well-known that, with competing risk data, nonparametric estimation of the marginal and conditional distributions of the competing outcomes is generally infeasible (Tsiatis 1975).

Semi-competing risk data are related to two types of events, a terminal event and a non-terminal event, where a terminal event such as death is always observed but the non-terminal event (such as disease progression or treatment interruption) may be censored by the terminal event. Semi-competing risk data are therefore an enriched form of competing risk data in the sense that the non-terminal event is not a competing cause of the terminal event but not vice versa. This special structure allows estimation of the marginal distribution of the terminal event. Using copula models for the joint distribution of the two types of events, the marginal distribution of the non-terminal event can also be identified, and the correlation between the two events can be modeled explicitly (Fine, Jiang, Chappell 2011, Wang 2003, Lakhal, Rivest, Abdous 2008).

Semi-competing risk data are frequently encountered. A typical example is the illness-death situation (Fix and Neyman 1951, Sverdrup 1965, Xu, Kalbfleisch, and Tai 2010), where death can censor the observation of illness if it occurs prior to illness, but occurrence of the illness may not prevent further follow-up on death. There are other semi-competing risk examples that do not fit into the illness-death framework. In AIDS studies, for example, time to first virologic failure and treatment discontinuation can be considered as the non-terminal and terminal events respectively (Jiang, Fine, Kosorok, and Chappell 2005). In oncology, they can be times to local and distant recurrences respectively (Dignam, Wieand, Rathouz 2007). Other interesting examples exist in biomarker studies (Ghosh 2009, Day, Bryant, and Lefkopolou 1997).

In this article, we consider the case when it is difficult or infeasible to ascertain all terminal events, especially in large cohort studies. This is the case of our motivating example, a cohort study of 8,977 adults who were enrolled between January 1, 2005 and

January 31, 2007 in the Academic Model Providing Access To Healthcare (AMPATH) program (Mamlin, Kimaiyo, Nyandiko, et al. 2004). AMPATH, the Nobel Peace Prize-nominated HIV care and treatment program, is a partnership between the Indiana University and Moi University Schools of Medicine in western Kenya. A key component in the evaluation of the effectiveness of this program is the marginal survival as well as the conditional survival (on being retained in care) distributions for patients under care. However, estimation of the distributions are severely complicated by the fact that there is a very high rate of patient loss to follow-up (Wools-Kaloustian et al., 2008). Moreover, there was evidence that the individuals lost to follow-up were generally sicker than those who remained on observation (An, Frangakis, Musick, and Yiannoutsos 2008, Yiannoutsos, An, Frangakis et al. 2008). These factors have the potential of introducing a significant bias if estimation is based only on data derived from patients under care.

Fortunately, AMPATH has instituted a major campaign to locate as many of the patients who are lost to follow-up as possible. The AMPATH patient outreach program uses location information available on all patients in an effort to ascertain the whereabouts of missing patients and attempt to persuade them to return to care. In the process, the program records, among other information, the vital status of all patients sought and successfully located. Using outreach data and theory on double sampling (Frangakis and Rubin 2001), estimation of the marginal survival distribution is possible (An et al. 2008). In this article, by viewing dropout as the non-terminal event and death as the terminal event, we revisit this estimation and consider more efficient estimators within the semi-competing risk data framework. In addition, we consider estimation of the conditional survival distributions and of the marginal distributions of the dropout in rural and urban areas that is informative for future policy making in providing care. Note that our considerations are applicable to the usual semi-competing risk situation (i.e. with 100% outreach), which is a special case of our setting.

The rest of the article is organized as follows. In Section 2 we consider estimation for the dependency between the terminal and non-terminal events using copula mod-

els. In Section 3 we discuss estimation of marginal distributions for non-terminal and terminal events. In particular, we investigate how to improve the efficiency of marginal distribution estimation for the terminal event by using the non-terminal event information. In Section 4, we consider estimation of the conditional survival distribution for the terminal event, conditioning on the non-terminal event. In Section 5, we consider model selection among possible models. In Section 6 we assess the performance of our revised copula estimators through simulations, using realistic parameters from our experience in this setting and perform a reanalysis of the AMPATH database (An et al. 2008 and Yiannoutsos et al. 2008). We conclude with a brief discussion of the implications of this methodological development in Section 7.

2 Modeling the dependency between X and Y

Let X be the time to the non-terminal event, Y the time to the terminal event, and C the administrative censoring time. In semi-competing risk data, X can be censored by Y if $Y < X$, but is observable if $Y \geq X$. However both X and Y can be censored by C. Therefore the observable quantities are: $R = Y \wedge C$, $\delta_R = 1(Y < C)$, $S = X \wedge Y \wedge C = X \wedge R$, and $\delta_S = 1(X < R)$ where $\wedge$ is the minimum operator. Now suppose that, among all subjects with $\delta_S = 1$, we only observe R in a subset. Therefore, we have an indicator variable η for each subject such that R is only observed if $\eta = 1$. We assume that $\pi \triangleq P(\eta = 1) = \delta_S p + (1 - \delta_S)$ where p is the proportion of dropouts who have been successfully located. Our objective is to estimate the marginal distributions $F_X(x) = P(X > x)$ and $F_Y(y) = P(Y > y)$, and the conditional survival distributions such as $P(Y > y|X = x, Y > t)$ and $P(Y > y|X > x, Y > t)$ for $y > t > x$.

We assume the following copula model for the joint survival distribution of X and Y (Oakes 1989):

$$F(x, y) = P(X > x, Y > y) = C_\alpha\{F_X(x), F_Y(y)\}$$

We are particularly interested in a class of copulas indexed by a single parameter α. A possible choice for C_α is the well known Archimedean copula with generator ϕ_α (Fine

et al. 2001, Wang 2003),

$$C_\alpha(u,v) = \phi_\alpha^{-1}\{\phi_\alpha(u) + \phi_\alpha(v)\}, \quad 0 \le u, v \le 1$$

Popular choices of ϕ_α include the Clayton copula (Clayton 1978): $\phi_\alpha(t) = t^{1-\alpha} - 1$ and the Gumbel copula: $\phi_\alpha(t) = (-\log t)^\alpha$, $\alpha \ge 1$.

A quantity which is useful in the effort to explicitly model the dependency between the terminal and non-terminal event is the cross-ratio function (Oakes 1989):

$$\theta_\alpha^*(x,y) = \frac{P\{(X_i - X_j)(Y_i - Y_j) > 0 | X_i \wedge X_j = x, Y_i \wedge Y_j = y\}}{P\{(X_i - X_j)(Y_i - Y_j) < 0 | X_i \wedge X_j = x, Y_i \wedge Y_j = y\}}. \tag{1}$$

This function is related to Kendall's tau (Wang 2003, Lakhal et al. 2008). Under the Archimedean copulas, θ_α depends on x and y only through their joint survival distribution $F(x,y)$ and takes the form

$$\theta_\alpha^*(x,y) = -F(x,y)\frac{\phi_\alpha''(F(x,y))}{\phi_\alpha'(F(x,y))} \triangleq \theta_\alpha(F(x,y)), \tag{2}$$

where $\theta_\alpha(v) = -v\phi_\alpha''(v)/\phi_\alpha'(v)$ for any v.

Utilizing (2), Lakhal et al. (2008) proposed the following estimating equation for general Archimedean copulas:

$$\Psi_n(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \left\{ \Delta_{ij} - \frac{\theta_\alpha\{\hat{F}(\tilde{S}_{ij}, \tilde{R}_{ij})\}}{1 + \theta_\alpha\{\hat{F}(\tilde{S}_{ij}, \tilde{R}_{ij})\}} \right\} \tag{3}$$

where $\tilde{S}_{ij} = S_i \wedge S_j$, $\tilde{R}_{ij} = R_i \wedge R_j$, and $\tilde{C}_{ij} = C_i \wedge C_j$; $\Delta_{ij} = 1\{(S_i - S_j)(R_i - R_j) > 0\}$; $Z_{ij} = 1(\tilde{S}_{ij} < \tilde{R}_{ij} < \tilde{C}_{ij})$; and W is a weighting function. Because

$$F(x,y) = \frac{P(S > x, R > y)}{P(C > y)}, \tag{4}$$

$F(x,y)$ is estimated by

$$\hat{F}(x,y) = \frac{n^{-1}\sum_{i=1}^n 1(S_i > x, R_i > y)}{\hat{G}(y)}$$

with

$$\hat{G}(y) = \prod_{i:C_i \le y} \left\{ 1 - \frac{1 - \Delta_{Ri}}{\sum_{k=1}^n 1(R_k \ge y)} \right\}$$

being the Kaplan-Meier estimator for the censoring distribution. Z_{ij} is needed in (3) because Δ_{ij} is evaluable only when $Z_{ij} = 1$ (Lakhal et al. 2008, Fine et al. 2001) and in this case Δ_{ij} equals the concordance between (X_i, Y_i) and (X_j, Y_j), that is, $1\{(X_i - X_j)(Y_i - Y_j) > 0\}$. The weighting function may be taken as 1 but a preferred form is given by (Fine et al. 2001)

$$W_{a,b}(x, y) = \frac{n}{\sum_{i=1}^{n} 1\{S_i \geq a \wedge x, R_i \geq b \wedge y\}} \tag{5}$$

where a and b are constants and may be selected to down weight for 'large' x and y.

In our case, because for subjects with $S_i < R_i$, R_i is observable only when $\eta_i = 1$, we propose the following estimating function when double sampling data are available:

$$\Psi_n^{\pi}(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^{\pi}(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \left\{ \Delta_{ij} - \frac{\theta_\alpha\{\hat{F}^{\pi}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}}{1 + \theta_\alpha\{\hat{F}^{\pi}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}} \right\}. \tag{6}$$

We can take the weighting function W^{π} as 1, or corresponding to (5), use

$$W_{a,b}^{\pi}(x, y) = \frac{n}{\sum_{i=1}^{n} \frac{\eta_i}{\pi_i} 1\{S_i \geq a \wedge x, R_i \geq b \wedge y\}} \tag{7}$$

$\hat{F}^{\pi}(x, y)$ estimates $F(x, y)$ and again take the following form because of (4),

$$\hat{F}^{\pi}(x, y) = \frac{\hat{P}(S > x, R > y)}{\hat{P}(C > y)} \tag{8}$$

When double sampled data are available, various estimators can be used for the numerator and denominator in (8). For $\hat{P}(S > x, R > y)$, we can use

$$\hat{P}^{\pi,1}(S > x, R > y) = n^{-1} \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} 1(S_i > x, R_i > y) \tag{9}$$

which inversely weighs the observations because of double sampling. However, an alternative choice is

$$\hat{P}^{\pi,2}(S > x, R > y) = n^{-1} \sum_{i=1}^{n} \left\{ 1(S_i > y) + \frac{\eta_i}{\pi_i} 1(S_i > x, S_i <= y, R_i > y) \right\} \tag{10}$$

where inverse weighting is only used for the subjects that need double sampling to ascertain the indicator $1(S_i > x, R_i > y)$. Because (10) utilizes more information, we use it in this article.

There are also two choices to estimate the censoring distribution in (8) under the double sampled data. One is

$$\hat{G}^{\pi}(y) = \prod_{i:C_i \le y} \left\{ 1 - \frac{\eta_i}{\pi_i} \frac{1 - \Delta_{Ri}}{\sum_{k=1}^{n} \frac{\eta_k}{\pi_k} 1(R_k \ge y)} \right\},$$

which utilizes inverse weighting. The other is

$$\tilde{G}(y) = \prod_{i:C_i \le y} \left\{ 1 - \frac{(1 - \Delta_{Si})(1 - \Delta_{Ri})}{\sum_{k=1}^{n} 1(S_k \ge y)} \right\}.$$

which uses $X \wedge Y$ as a censoring variable for C and does not involve inverse weighting. We use $G\tilde{(y)}$ in (8) because greater efficiency gains were observed in simulations.

By recognizing that the weighted estimator is related to U-statistics, we derive the asymptotic properties of the resulting estimator $\hat{\alpha}^{\pi}$ from (6) in the Appendix. Because the asymptotic variance is complicated, we use a bootstrap procedure to obtain standard error estimates as in Lakhal et al (2008).

Finally, we note that for the Clayton copula, $\theta^*_{\alpha} = \alpha$ is not related to either x or y and $P(\Delta_{ij} = 1) = \frac{\alpha}{1+\alpha}$ for any $i \neq j$ (Oakes 1989, Fine et al. 2001). Then it is easily seen that an explicit solution for α exists from (3):

$$\hat{\alpha}^{\pi} = \frac{\sum_{i<j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^{\pi}(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Delta_{ij}}{\sum_{i<j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^{\pi}(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} (1 - \Delta_{ij})}.$$

When $\pi \equiv 1$ in the semi-competing risk data, this reduces to the estimator from Fine et al. (2001).

3 Estimation of marginal distributions

3.1 Estimation of the marginal survival distribution of Y

Adopting the counting process notation (Kalbfleisch and Prentice 2002), we estimate the cumulative hazard function of Y by

$$\hat{\Lambda}_Y^{\pi,1}(t) = \int_0^t \frac{\sum_{i=1}^{n} (1 - \delta_{Si}) \mathrm{d}N_i(u) + \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} \delta_{Si} \mathrm{d}N_i(u)}{\sum_{i=1}^{n} (1 - \delta_{Si}) 1(R_i \ge u) + \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} \delta_{Si} 1(R_i \ge u)}$$

where $N_i(u) = 1(Y_i \leq u, C_i > Y_i)$. This estimator was proposed in Frangakis and Rubin (2001). The corresponding survival function estimate is then

$$\hat{F}_Y^{\pi,1}(t) = \exp\left\{-\sum_{i=1}^{n} \hat{\Lambda}_{Yi}^{\pi,1}(t)\right\}. \tag{11}$$

An alternative way is to estimate the cumulative hazard function of Y by

$$\hat{\Lambda}_Y^{\pi,2}(t) = \int_0^t \frac{\sum_{i=1}^n (1-\delta_{Si})\mathrm{d}N_i(u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i}\delta_{Si}\mathrm{d}N_i(u)}{\sum_{i=1}^n (1-\delta_{Si})1(R_i \geq u) + \sum_{i=1}^n \delta_{Si}1(S_i \geq u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i}\delta_{Si}1(S_i < u)1(R_i \geq u)}.$$

This estimator has been mentioned in Robins, Rotnitzky, and Bonetti (2001). Notice that the calculation of the at-risk set at any time u uses the similar idea as in the estimation of $\hat{P}^{\pi,2}(S > x, R > y)$ in (10) where inverse weighting is only used for the subjects that need double sampling to ascertain the at risk status. The corresponding survival function estimate is then

$$\hat{F}_Y^{\pi,2}(t) = \exp\left\{-\sum_{i=1}^{n} \hat{\Lambda}_{Yi}^{\pi,2}(t)\right\}. \tag{12}$$

Asymptotics for both (11) and (12) can also be established in a similar fashion as in Yu and Nan (2010).

Neither approach utilizes the correlation between X and Y. We therefore consider a model-based estimator. In particular, by partitioning $P(Y > t)$ as

$$P(Y > t) = P(Y \wedge C > t) + P(Y > t, C \leq t, X < C) + P(Y > t, C \leq t, X > C),$$

we propose the following estimate

$$\begin{aligned}\hat{F}_Y^{\pi,M}(t) = \frac{1}{n}\sum_{i=1}^{n} 1(R_i > t)\big\{(1-\delta_{Si}) + \frac{\eta_i}{\pi_i}\delta_{Si}\big\} \\ + \frac{1}{n}\sum_{i=1}^{n} 1(R_i \leq t)(1-\delta_{Ri})\frac{\eta_i}{\pi_i}\delta_{Si}\hat{P}(Y > t | X = S_i, Y > R_i) \\ + \frac{1}{n}\sum_{i=1}^{n} 1(R_i \leq t)(1-\delta_{Ri})(1-\delta_{Si})\hat{P}(Y > t | X > S_i, Y > S_i).\end{aligned} \tag{13}$$

Here the $\hat{P}(Y > t | X = S_i, Y > R_i)$ and $\hat{P}(Y > t | X > S_i, Y > S_i)$ are conditional estimates of the corresponding probabilities. The estimates will be discussed in Section

4. To avoid a circular problem, we use $\hat{F}_Y^{\pi,2}$ in estimation of these conditional distributions (see Section 4) because $\hat{F}_Y^{\pi,2}$ is generally more efficient than $\hat{F}_Y^{\pi,1}$ (see Figure 1).

As we demonstrate in our numerical section, compared to (11) and (12), $\hat{F}_Y^{\pi,M}(t)$ improves efficiency noticeably when there is moderate to high correlation between X and Y. Such improvement is important especially when a large number of subjects experience X and double sampling for Y is costly (i.e., p is small).

3.2 Estimation of the marginal distribution of X

We adapt the copula graphic (CG) estimator for the marginal distribution $F_X(x)$ from Lakhal et al. (2008) in our double sampled data. The CG estimator utilizes the following relationship (Zheng & Klein 1995):

$$F_X(x) = \phi_\alpha^{-1}\{\phi_\alpha[F(x,x)] - \phi_\alpha[F_Y(x)]\}$$

where $F(x,x) = P(X \wedge Y > x) \triangleq F_{X\wedge Y}(x)$ and F_Y is the marginal distribution of Y. By replacing unknown parameters with estimates, the resulting CG estimator is

$$\hat{F}_X^{CG}(x) = \phi_{\hat{\alpha}}^{-1}\Big\{\sum_{S_i\le x,\delta_{Si}=1} \phi_{\hat{\alpha}}[\hat{F}_{X\wedge Y}(S_i)] - \phi_{\hat{\alpha}}[\hat{F}_{X\wedge Y}(S_i-)]\Big\}. \tag{14}$$

Here $\hat{F}_{X\wedge Y}$ is the Kaplan-Meier estimator for $X \wedge Y$ which needs no inverse weighting because observation of $X \wedge Y$ is unaffected by double sampling. Note that because F_Y is discrete and jumps only at observed failure times of Y, the term involving F_Y disappears in (14) (Rivest and Wells 2001, Lakhal et al. 2008). The CG estimator is therefore a discrete function that jumps only at observed failure times of X.

In the presence of double sampling, because S_i are always observed and therefore are used in the same way to estimate $F(x,x)$ and because F_Y is not involved in (14), the CG estimator has exactly the same format as in (14), i.e.,

$$\hat{F}_X^{\pi,CG}(x) = \phi_{\hat{\alpha}^\pi}^{-1}\Big\{\sum_{S_i\le x,\delta_{Si}=1} \phi_{\hat{\alpha}^\pi}[\hat{F}_{X\wedge Y}(S_i)] - \phi_{\hat{\alpha}^\pi}[\hat{F}_{X\wedge Y}(S_i-)]\Big\}. \tag{15}$$

The asymptotic distribution of this estimator is therefore exactly the same as in the complete ascertainment case, except that we replace $\hat{\alpha}$ with $\hat{\alpha}^\pi$ given as a solution of (6), taking into consideration the asymptotic normality of $\hat{\alpha}^\pi$ from the Appendix.

4 Estimation of Conditional Distributions

The conditional survival function of a patient that has not dropped out at time x is $P(Y > y|X > x; Y > x)$. Note here that the ability to evaluate this conditional probability is particularly relevant in our context, since it is useful in directly addressing the question of what the survival distribution is for a patient who is retained in care up to time $x > x^*$, where x^* is a clinically meaningful threshold (e.g., three or six months from start of antiretroviral therapy, a critical period in the care and treatment of HIV-infected patients initiating therapy).

Just like in the estimation of the marginal distribution of Y, two different estimators can be used for $P(Y > y|X > x; Y > x) = F(x, y)/F(x, x)$. One is nonparametric:

$$\hat{P}^{\pi}(Y > y|X > x; Y > x) = \frac{\hat{F}^{\pi}(x, y)}{\hat{F}_{X\wedge Y}(x)}$$

where $\hat{F}^{\pi}$ is defined in (8). The other is model based:

$$\hat{P}^{M}(Y > y|X > x; Y > x) = \frac{C_{\hat{\alpha}^{\pi}}\{\hat{F}_X^{\pi,CG}(x), \hat{F}_Y^{\pi,2}(y)\}}{C_{\hat{\alpha}^{\pi}}\{\hat{F}_X^{\pi,CG}(x), \hat{F}_Y^{\pi,2}(x)\}} \tag{16}$$

with $\hat{F}^{\pi,CG}$ from (15) and $\hat{F}_Y^{\pi,2}$ from (12). Of course $\hat{F}_Y^{\pi,1}$ from (11) can also be used in (16) with some loss of efficiency.

The conditional survival function of a patient who drops out at time x and is alive at x is $P(Y > y|X = x; Y > x)$ for $y > x$. It satisfies (Lakhal et al. 2008)

$$P(Y > y|X = x; Y > x) = \frac{\phi'_{\alpha}\{F(x, x)\}}{\phi'_{\alpha}\{F(x, y)\}} = \frac{\phi'_{\alpha}\{F_{X\wedge Y}(x)\}}{\phi'_{\alpha}\{F(x, y)\}}.$$

with $\phi'_{\alpha}(t)$ the derivative of $\phi_{\alpha}(t)$ with respect to t. We can then obtain an estimator by plugging in estimators for α, $F_{X\wedge Y}(x)$, and $F(x, y)$. Similar to estimation of $\hat{P}^{M}(Y > y|X > x; Y > x)$, we can use a nonparametric estimator,

$$\hat{P}^{\pi}(Y > y|X = x; Y > x) = \frac{\phi'_{\hat{\alpha}^{\pi}}\left[\hat{F}^{\pi}(x, y)\right]}{\phi'_{\hat{\alpha}^{\pi}}\left[\hat{F}_{X\wedge Y}(x)\right]}$$

or a model based estimator,

$$\hat{P}^{M}(Y > y|X = x; Y > x) = \frac{\phi'_{\hat{\alpha}^{\pi}}\left[C_{\hat{\alpha}^{\pi}}\{\hat{F}_X^{\pi,CG}(x), \hat{F}_Y^{\pi,2}(y)\}\right]}{\phi'_{\hat{\alpha}^{\pi}}\left[C_{\hat{\alpha}^{\pi}}\{\hat{F}_X^{\pi,CG}(x), \hat{F}_Y^{\pi,2}(x)\}\right]} \tag{17}$$

The performance of all these semi-parametric estimators of the conditional survival distributions is explored through simulations.

5 Model selection

The idea is to choose a quantity that is both estimable by the data (nonparametrically) and by the proposed models. Among the proposed models, the model which produces the closest estimate to the quantity is deemed the best. Such quantity should not be used already in the estimating procedure and sensitive to different models. One such quantity is $F(x,y;\delta_S=\delta_R=1) \triangleq P(X>x, Y>y, \Delta_S=1, \Delta_R=1)$, which is estimable nonparametricly by

$$\hat{F}(x,y;\delta_S=\delta_R=1) \triangleq n^{-1}\sum_{i=1}^{n}\frac{\eta_i}{\pi_i}1(S_i>x, R_i>y, \Delta_{Si}=1, \Delta_{Ri}=1).$$

Under an Archimedean copula model,

$$F(x,y;\delta_S=\delta_R=1) = \int_y^{\infty}\left\{\frac{\phi'_\alpha\big[F_Y(t)\big]f_Y(t)}{\phi'_\alpha\big[F(x,t)\big]} - \frac{\phi'_\alpha\big[F_Y(t)\big]f_Y(t)}{\phi'_\alpha\big[F(t,t)\big]}\right\}G(t)\mathrm{d}t$$

Therefore it can also be estimated by

$$\tilde{F}(x,y;\delta_S=\delta_R=1) = \int_y^{\infty}\left\{\frac{\phi'_{\hat{\alpha}^\pi}\big[\hat{F}_Y^{\pi,2}(t)\big]}{\phi'_{\hat{\alpha}^\pi}\big[\hat{F}^\pi(x,t)\big]} - \frac{\phi'_{\hat{\alpha}^\pi}\big[\hat{F}_Y^{\pi,2}(t)\big]}{\phi'_{\hat{\alpha}^\pi}\big[\hat{F}^\pi(t,t)\big]}\right\}\hat{G}(t)\mathrm{d}\hat{F}_Y^{\pi,2}(t)$$

The maximum distance, $\mathcal{D} \triangleq max_{0<x<y}|\hat{F}(x,y;\delta_S=\delta_R=1) - \tilde{F}(x,y;\delta_S=\delta_R=1)|$ is then used as a criterion to choose different models. In other words, among possible AC models, the model with the smallest difference measure $\mathcal{D}$ is chosen. This builds on the model checking method of Hsieh, Wang and Adam (2008).

Of course, it is likely that no model among the proposed ones provides adequate fit to the data. In this case, calibration by p-values may be desirable for model selection purposes (Hsieh et al. 2008). A similar bootstrap approach will be taken to approximate the null distribution of the difference measure. Specifically, bootstrap samples of (X,Y) will be generated from the estimated copula model and C from $\hat{G}$. The bootstrap samples will then be censored according to semi-competing risk framework. The model fitting measures are then calculated for all the bootstrap samples to create the null distribution.

6 Numerical examples

In this section we assess the performance of various proposed estimators and the model selection procedure. We also analyze the data from the AMPATH study.

6.1 Simulations

In generating data, the marginal distribution of Y is taken to be unit exponential and the marginal distribution of X exponential with rate 0.8. This leads to about 40 percent dropout. This is close to the rate of 39.3% reported in An, Yiannoutsos, Frangakis et al. (2008). The censoring distribution is uniform within the interval $(0, 4)$. To save space, we present results using the Clayton and Gumbel copulae with various α for the correlation between X and Y. We used the bootstrap procedure for standard error estimation in all analyses.

In the following sections, we present results for estimation of the correlation between X and Y in Table 1; for the marginal distributions of X and Y in Table 2; and for the conditional distributions of Y given X in Table 3. A total of 400 data sets were simulated in each scenario with various correlations between X and Y. The sample sizes were set to be 1000 and 3000 in each setting. The double sampling proportion is set to be 0.2.

In Table 1, we see that our estimation procedure performs satisfactorily in all cases. The standard error decreases with larger sample size $n = 3000$ as compared with $n = 1000$. In Table 2, we try to estimate the marginal distribution of X at one year (i.e., $t = 1$) and the marginal distribution of Y at 2 years ($t = 2$). The CG estimates are all close to the true value $\exp(-1) = 0.368$ and the estimates for $F_Y(2)$ are all close to the true value $\exp(-0.8 * 2) = 0.202$. Among the standard error estimates of the three estimators of F_Y, we see that $\hat{F}_Y^{\pi,1}$ is the most inefficient and $\hat{F}_Y^M$ the most efficient among the three. The improvement of efficiency is actually related to the strength of correlation between X and Y. Figure 1 illustrates the efficiency comparison among these marginal distribution estimators under the two copulae. Here the x-axis represents the correlation between X and Y (instead of α values in the copulae). The results are based on the same setting as in Table 1 but with 50,000 simulated data.

We see that $\hat{F}_Y^{\pi,2}$ is in general more efficient than $\hat{F}_Y^{\pi,1}$ except in the case of very high correlations. When the correlation is small, model based estimator can be inefficient. However the efficiency improves dramatically as correlation increases. This implies that the model-based estimator may be preferred if there is moderate to high correlations.

Table 3 gives results for estimation of conditional distributions $F_{1.5|.75} \triangleq P(Y > 1.5|X = 0.75, Y > 0.75)$ and $F_{1.5|.75+} \triangleq P(Y > 1.5|X > 0.75, Y > 0.75)$. Again the results are satisfactory. Finally, we conducted model selection procedures under the same setting with results listed in Table 4. We can see that correct models were selected with very high percentage in all cases. The correct selection percentages are higher with the larger sample size as expected.

6.2 The AMPATH study

The study included 8,977 adults coming from both urban and rural clinics in western Kenya between January 1, 2005 and January 31, 2007. There is a very high rate of patient loss to follow-up (3,528 dropouts). In the outreach program, 621 were double sampled for further follow-up. The initial goal of AMPATH was to establish an HIV care system to serve the needs of both urban and rural patients and to assess the barriers to and outcomes of antiretroviral therapy. It is important to explore differences in dropout patterns and patient survival between the rural and urban areas. For the urban area, there were 6561 subjects among whom 38% dropped out. The outreach rate is 18%. The maximum follow up is 761 days with a total of 136 deaths observed. For the rural area, there were 2416 subjects among whom 43% dropped out. The outreach rate is 23%. The maximum follow up is 761 days with a total of 86 deaths observed.

The analysis results listed in Table 5 are based on the Clayton copula which was selected based on the difference measure proposed in Section 5. The marginal distributions for dropout and survival in rural and urban clinics are plotted in Figure 2. The conditional survival distributions in rural and urban clinics are plotted in Figure 3.

In Table 5, the estimate of the α index of the Clayton copula (s.e.) was 2.865 (0.481) among patients enrolled in the urban clinic and 1.852 (0.419) for rural clinics.

These estimates correspond to correlations of 0.59 and 0.48, which in turn suggests that patient dropout is non-ignorable. We can also see that the rural clinics appear to have inferior outcomes in several aspects (Table 5). Among the statistical significant differences, the non-dropout rate at 1 year is 59.8% compared with 62.4% in the urban area. Patients retained in care for 3 or 6 months, also have shorter survival when attending rural clinics versus receiving care from the referral hospital. From the marginal distribution curve of dropout, we find that the dropout rates are similar in the early period (up to about 180 days) and then the dropout rate increases in the rural area (Figure 2).

The conditional survival probabilities are seen to increase noticeably when subjects are kept in care longer. In Figure 3, we see a similar increase in conditional survival for both urban and rural clinics comparing dropout at 3 months with 6 months, or comparing survival given retention in care > 3 months with > 6 months.

7 Discussion

In this article, we considered various estimation aspects and model selection for mortality and dropout in the context of a large HIV care and treatment program in western Kenya. We developed novel statistical methods in a semi-competing risk framework where we were able to ascertain only a fraction of the terminal events. By utilizing the correlation between the non-terminal and terminal events, we demonstrated that improved estimation for the marginal and conditional distribution of the terminal event is feasible as is the marginal distribution of the non-terminal event.

In addition to proving the asymptotic properties of the proposed estimators (Appendix), we performed simulations to assess their performance under scenarios closely resembling the real-life settings where these methods will be applied. In all cases, the performance of these estimators was excellent. This has a number of important implications for the practical application of our methods. It means that program evaluation by assessment of HIV care and treatment program effectiveness metrics such as mortality, patient retention, treatment interruption or disease progression rates, all falling under our semi-competing risk framework, can be accomplished by locating only

a relatively small subset of patients who are lost. This ensures the viability and sustainability of ongoing program evaluation efforts in this setting as most programs in low and middle-income countries in sub-Saharan Africa, Asia and Latin America are so large and the loss to follow-up problem so pronounced (Rosen, Fox, and Gill 2007) as to make it infeasible to follow-up all patients who discontinue treatment.

Our re-analysis of the AMPATH data (An, et al., 2008, Yiannoutsos, et al., 2008) showed the advantages of being able to estimate conditional survival distributions for the terminal event conditional on loss-to-follow-up (the non-terminal event). We considered three and six months, two important landmarks from start of therapy, when the hazard of mortality is highest (Yiannoutsos, 2009). Conditioning on being retained on treatment (and thus being alive) at 3 and 6 months after therapy produced estimates of mortality (Figure 3 right panel), which were much lower than overall (marginal) mortality (Figure 2 right panel). In addition, patient retention for at least six months is associated with substantial improvement of mortality (both early and late) compared to retention at least for three months (Figure 3). These results have important policy implications about interventions which target patients early after initiating antiretroviral therapy. AMPATH for example, has instituted an intense follow-up of patients during the first three months after treatment initiation. These analyses provide strong supportive evidence for such interventions.

While these analyses are intended more as a proof of concept rather than an exhaustive analysis of all possible explanatory factors of dropout and mortality patterns in our setting, these first results underline some important issues and further emphasize the practical utility of these methods. The ability to explicitly model the possible correlation between X and Y allows both the assessment of the strength of the association as well as the comparison of these quantities between groups. In our re-analysis of the AMPATH data, we see a possibly different pattern in the urban versus the rural settings with dropout being associated less strongly with mortality in the rural setting. These observations are important because they underscore the necessity of methods which do not rely on independence assumptions between X and Y and the absolute critical nature of outreach and vital status ascertainment among lost patients. In ad-

dition, these findings could be used to tailor contextually appropriate interventions depending on the setting (rural versus urban).

While our motivating example was focused on issues arising in the evaluation of large care and treatment programs recently proliferating in low and middle-income countries, the applicability of the proposed methods is quite broad. The extent of the application of this methodology is related both to the number of issues in this context as well as any other chronic disease intervention requiring patient retention and long-term adherence to treatment or prevention interventions.

ACKNOWLEDGEMENT

Dr. Yiannoutsos' work was supported by grant number AI-069911 from the National Institute of Allergies and Infectious Diseases (NIAID), for the East Africa Regional IeDEA Consortium. The data were collected through a grant to the USAID-AMPATH Partnership from the United States Agency for International Development as part of the President's Emergency Plan for AIDS Relief (PEPFAR). The funders had no role in study design, interpretation of results, or decision to publish.

REFERENCES

An M-W, Frangakis CE, Musick BS, Yiannoutsos CT. The need for double-sampling designs in survival studies: an application to monitor PEPFAR. *Biometrics* 2008; **65**: 301–306.

Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–151.

Day R, Bryant J, Lefkopolou M. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika* 1997; **84**: 45–56.

Dignam JJ, Wieand K, and Rathouz PJ. A missing data approach to semi-competing risks problems. *Statistics in Medicine* 2007; **26**:837–56.

Fine JP, Jiang H, Chappell R. On semicompeting risks. *Biometrika* 2001; **88**: 907–919.

Fix E and Neyman J. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology* 1951; **23**:205–41.

Frangakis CE, Rubin DB. Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics* 2001; **57**: 333–342.

Ghosh D. Semiparametirc inferences for association with semi-competing risks data. *Statistics in Medicine* 2006; **25**:2059–70.

Jiang H, Fine JP, Kosorok MR, and Chappell R. Pseudo self-consistent estimation of a copula model with informative censoring. *Scandinavian Journal of Statistics* 2005; **3**2:1–20.

Klein JP, Moeschberger ML. *Survival analysis: Techniques for censored and truncated data* 1997. New York: Springer.

Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data*, Second Ed. 2002. New York, John Wiley & Sons, Inc.

Lakhal L, Rivest L-P, Abdous B. Estimating survival and association in a semicompeting risks model. *Biometrics* 2008; **64**: 180–188.

Mamlin J, Kimaiyo S, Nyandiko W, Tierney W, Einterz R. Academic institutions linking access to treatment and prevention: Case study. *In: Perspectives and Practice in Antiretroviral Treatment*. Geneva: World Health Organization; 2004.

Oakes D. Bivariate survival models induced by frailties. *J Amer Stat Assoc* 1989; **84**: 487–493

Rivest LP and Wells MT. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J Multivar Analysis* 2001; **79**: 138–155.

Robins JM, Rotnitzky A, and Bonetti M. Discussion of the Frangakis and Rubin article, "Addressing an idiosyncrasy in estimating survival curves using double sampling

in the presence of self-selected right censoring." *Biometrics* 2001; **57**:343–347.

Rosen S, Fox MP, Gill CJ. Patient retention in antiretroviral therapy programs in sub-Saharan Africa: A systematic review. *PLoS Medicine* 2007; **4**: e298.

Sverdrup E. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. Skand. Aktuarietietidskr 1965; 48:184–211.

Tsiatis A. A nonidentifiability aspect of the problem of competing risks.*Proc Nat Acad Sci USA* 1975; **72**: 20–22.

Wang W. Estimating the association parameter for copula models under dependent censoring. *J Royal Stat Soc B* 2003; **65**: 257–273.

Xu J, Kalbfleisch JD, Tai B. Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* 2010; **66**:716–725.

Yiannoutsos CT, An M-W, Frangakis CE, Musick BS, Braitstein P, et al. Sampling-based approaches to improve estimation of mortality among patient dropouts: Experience from a large PEPFAR-funded program in western Kenya. *PLoS ONE* 2008; **3**: e3843.

Yiannoutsos CT. Modeling AIDS survival after initiation of antiretroviral treatment by Weibull models with changepoints. *J Int AIDS Soc* 2009; **12**: 9.

Yu M and Nan B. Regression calibration in semiparametric accelerated failure time models. *Biometrics* 2010; **66**; 405–414.

Zheng M, Klein J. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**: 127–138.

Appendix

We show the consistency and asymptotic distribution of $\hat{\alpha}^{\pi}$, the estimator of the dependency between X and Y in the context of semi-competing risks with incomplete vital status ascertainment. Assume that the outreach probability $p > 0$, then similar to the arguments in the web appendix of Lakhal et al. (2008), for any $\epsilon > 0$, all comparable pairs belong to a bounded set $\mathcal{S}_0 = \{(x, y) : \tau_0 > y \geq x > 0\}$ with a high probability $1-2\epsilon^2$, where τ_0 is a fixed number such that $0 < F_X(\tau_0)G(\tau_0) < \epsilon, 0 < F_Y(\tau_0)G(\tau_0) < \epsilon$, and $0 < F(\tau_0, \tau_0)G(\tau_0)$. Then because $\hat{G}(y)$, $\hat{P}^{\pi,2}$ are consistent estimators of their limit in the set $\mathcal{S}_0$, $\hat{F}^{\pi}(x, y)$ is consistent for $F(x, y)$ in $\mathcal{S}_0$. This leads to the consistency of $\hat{\alpha}^{\pi}$ because (6) converges to its limit and is monotonic as a function of α for Archimedean copulae.

For the asymptotic distribution, consider a Taylor expansion of the function $\xi_\alpha(v) = \{1 + \theta_\alpha(v)\}^{-1}\theta_\alpha(v)$. The estimating function $\Psi_n(\alpha)$ is asymptotically equivalent to

$$\begin{aligned}\Psi_n(\alpha) \;\simeq\; \binom{n}{2}^{-1} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Big[\Delta_{ij} - \xi_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \\ - \xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\}\{\hat{\pi}(\tilde{X}_{ij}, \tilde{Y}_{ij}) - \pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \Big]\end{aligned}$$

Because

$$\begin{aligned}\hat{\pi}(x, y) - \pi(x, y) \quad &= n^{-1}\hat{G}(y)^{-1} \sum_{k=1}^{n} 1(S_k > x, R_k > y) - G(y)^{-1}P(S_k > x, R_k > y) \\ &= n^{-1}\{\hat{G}(y)^{-1} - G(y)^{-1}\} \sum_{k=1}^{n} 1(S_k > x, R_k > y) \\ &\quad + n^{-1}G(y)^{-1} \sum_{k=1}^{n} \{1(S_k > x, R_k > y) - P(S_k > x, R_k > y)\}\end{aligned}$$

we can break $\Psi_n(\alpha)$ into two parts $\Psi(\alpha) = A + B$ where,

$$\begin{aligned} A &= -\binom{n}{2}^{-1} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Big[\xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} n^{-1} \{\hat{G}(\tilde{Y}_{ij})^{-1} - G(\tilde{Y}_{ij})^{-1}\} \\ &\qquad \times \sum_{k=1}^{n} 1(S_k > \tilde{X}_{ij}, R_k > \tilde{Y}_{ij}) \Big] \\ &= -\binom{n}{2}^{-1} n^{-1} \sum_{k=1}^{n} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \{\hat{G}(\tilde{Y}_{ij})^{-1} - G(\tilde{Y}_{ij})^{-1}\} \\ &\qquad \times 1(S_k > \tilde{X}_{ij}, R_k > \tilde{Y}_{ij}) \end{aligned}$$

and

$$\begin{aligned} B &= \binom{n}{2}^{-1} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Big[\Delta_{ij} - \xi_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \\ &\qquad -\xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} n^{-1} G(\tilde{Y}_{ij})^{-1} \sum_{k=1}^{n} \{1(S_k > \tilde{X}_{ij}, R_k > \tilde{Y}_{ij}) - P(S_k > \tilde{X}_{ij}, R_k > \tilde{Y}_{ij})\} \Big] \\ &= \binom{n}{2}^{-1} n^{-1} \sum_{k=1}^{n} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Big[\Delta_{ij} - \xi_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \\ &\qquad -\xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \{G(y)^{-1} 1(S_k > \tilde{X}_{ij}, R_k > \tilde{Y}_{ij}) - \pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \Big]. \end{aligned}$$

The first part is asymptotically equivalent to

$$A \simeq -\binom{n}{2}^{-1} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \{\hat{G}^{-1}(\tilde{Y}_{ij}) - G^{-1}(\tilde{Y}_{ij})\} \pi(\tilde{X}_{ij}, \tilde{Y}_{ij}) / G(\tilde{Y}_{ij})$$

because $\{\hat{G}^{-1}(\tilde{Y}_{ij}) - G^{-1}(\tilde{Y}_{ij})\}$ can be written as an iid random sum (Gill 1980 p37).
The second part is asymptotically equivalent to

$$B \simeq -\binom{n}{2}^{-1} \frac{1}{n} \sum_{k=1}^{n} \sum_{i<j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Big\{ \frac{1(C_k \le t)(1 - \Delta_{ZXk})}{S_Z(C_k-)} - G(\tilde{Y}_{ij}) \Big\} \xi'_\alpha\{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})\} \frac{\pi(\tilde{X}_{ij}, \tilde{Y}_{ij})}{G(\tilde{Y}_{ij})}$$

This is a U-statistic with order 3. Asymptotic normality then directly follows from the theory of U-statistics.

Table 1: Simulation results for the estimation of α

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Clayton; $n = 1000$			
Mean Est	2.058	3.038	4.117
SE	.411	.536	.658
Emp SE	.410	.509	.660
95% CP	94.0	96.0	94.5
Clayton; $n = 3000$			
Mean Est	1.993	3.035	4.011
SE	.228	.298	.363
Emp SE	.237	.290	.354
95% CP	92.8	95.2	96.8
Gumbel; $n = 1000$			
Mean Est	2.023	2.935	4.001
SE	.237	.407	.605
Emp SE	.259	.418	.619
95% CP	91.8	92.8	93.2
Gumbel; $n = 3000$			
Mean Est	2.077	2.972	3.965
SE	.135	.234	.335
Emp SE	.135	.251	.328
95% CP	92.8	92.5	93.2

Table 2: Simulation results for the estimation of marginal distributions at times $t = 1$ (F_X) and $t = 2$ (F_Y). The target quantities are $F_X(1) = \exp(-1) = 0.368$ and $F_Y(2) = \exp(-0.8 \times 2) = 0.202$.

	$\alpha = 2$				$\alpha = 3$				$\alpha = 4$			
	$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$	$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$	$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$
Clayton; $n = 1000$												
Mean Est	.368	.204	.205	.203	.369	.206	.209	.205	.368	.206	.208	.206
SE	.021	.030	.029	.028	.020	.030	.029	.027	.018	.030	.030	.026
Emp SE	.021	.032	.030	.029	.019	.033	.031	.029	.018	.030	.029	.026
95% CP	95.5	93.5	93.5	93.5	96.0	91.8	92.2	91.8	95.0	95.2	94.2	96.0
Clayton; $n = 3000$												
Mean Est	.368	.204	.205	.204	.368	.203	.205	.203	.368	.203	.205	.203
SE	.012	.018	.017	.016	.011	.018	.017	.016	.011	.018	.017	.016
Emp SE	.012	.018	.018	.016	.011	.019	.017	.017	.011	.019	.017	.017
95% CP	94.5	96.0	93.8	95.5	92.8	93.2	95.5	92.8	92.8	93.2	95.5	92.8
Gumbel; $n = 1000$												
Mean Est	.372	.208	.209	.208	.372	.205	.206	.206	.369	.208	.209	.207
SE	.021	.032	.030	.031	.019	.031	.029	.029	.018	.031	.030	.028
Emp SE	.022	.034	.032	.033	.018	.032	.032	.030	.019	.031	.031	.028
95% CP	93.0	92.0	91.8	93.8	96.8	93.5	92.5	93.5	91.5	93.2	93.0	93.8
Gumbel; $n = 3000$												
Mean Est	.370	.204	.204	.204	.370	.203	.202	.202	.370	.203	.202	.202
SE	.012	.019	.018	.018	.011	.018	.017	.017	.011	.018	.017	.017
Emp SE	.013	.020	.019	.020	.011	.019	.017	.017	.011	.019	.017	.017
95% CP	93.0	93.0	93.8	94.0	93.0	93.8	95.5	94.0	93.0	93.8	95.5	94.0

Table 3: Simulation results for the estimation of the conditional distributions.

	$\alpha = 2$		$\alpha = 3$		$\alpha = 4$	
	$\hat{F}_{1.5\|.75}$	$\hat{F}_{1.5\|.75+}$	$\hat{F}_{1.5\|.75}$	$\hat{F}_{1.5\|.75+}$	$\hat{F}_{1.5\|.75}$	$\hat{F}_{1.5\|.75+}$
Clayton; $n = 1000$						
Mean Est	.207	.587	.127	.585	.072	.576
SE	.070	.062	.063	.065	.051	.068
Emp SE	.074	.066	.065	.070	.048	.069
95% CP	92.0	92.0	94.8	90.2	95.0	92.2
Clayton; $n = 3000$						
Mean Est	.207	.588	.116	.581	.064	.569
SE	.041	.036	.035	.039	.026	.041
Emp SE	.044	.038	.034	.038	.028	.044
95% CP	92.0	93.0	94.2	93.5	94.0	92.2
Gumbel; $n = 1000$						
Mean Est	.363	.533	.246	.545	.162	.556
SE	.067	.060	.064	.062	.058	.065
Emp SE	.069	.066	.067	.067	.058	.068
95% CP	93.5	90.5	91.5	91.5	95.0	92.8
Gumbel; $n = 3000$						
Mean Est	.355	.526	.233	.540	.150	.545
SE	.040	.036	.038	.038	.032	.039
Emp SE	.043	.039	.039	.038	.032	.040
95% CP	93.0	94.0	94.8	95.5	95.0	94.5

Table 4: Correct model selection percentage results based on 400 simulated data sets.

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
$n = 1000$			
Clayton	96	95	81
Gumbel	93	98	98
$n = 3000$			
Clayton	100	99	89
Gumbel	100	100	100

Table 5: Data analysis results for the AMPATH study.

	Urban		Rural		P-value
Estimand	Est	SE	Est	SE	(Urban vs. Rural)
α	2.865	.481	1.852	.419	.029
$P(X > 1yr)$	.624	.007	.598	.013	.004
$P(Y > 1yr)$	.915	.014	.892	.024	.183
$P(Y > 1yr \mid X = 3mo, Y > 3mo)$	.864	.018	.846	.028	.395
$P(Y > 1yr \mid X > 3mo, Y > 3mo)$	.958	.006	.941	.010	.020
$P(Y > 1yr \mid X = 6mo, Y > 6mo)$	.949	.014	.930	.021	.240
$P(Y > 1yr \mid X > 6mo, Y > 6mo)$	.985	.004	.974	.008	.041

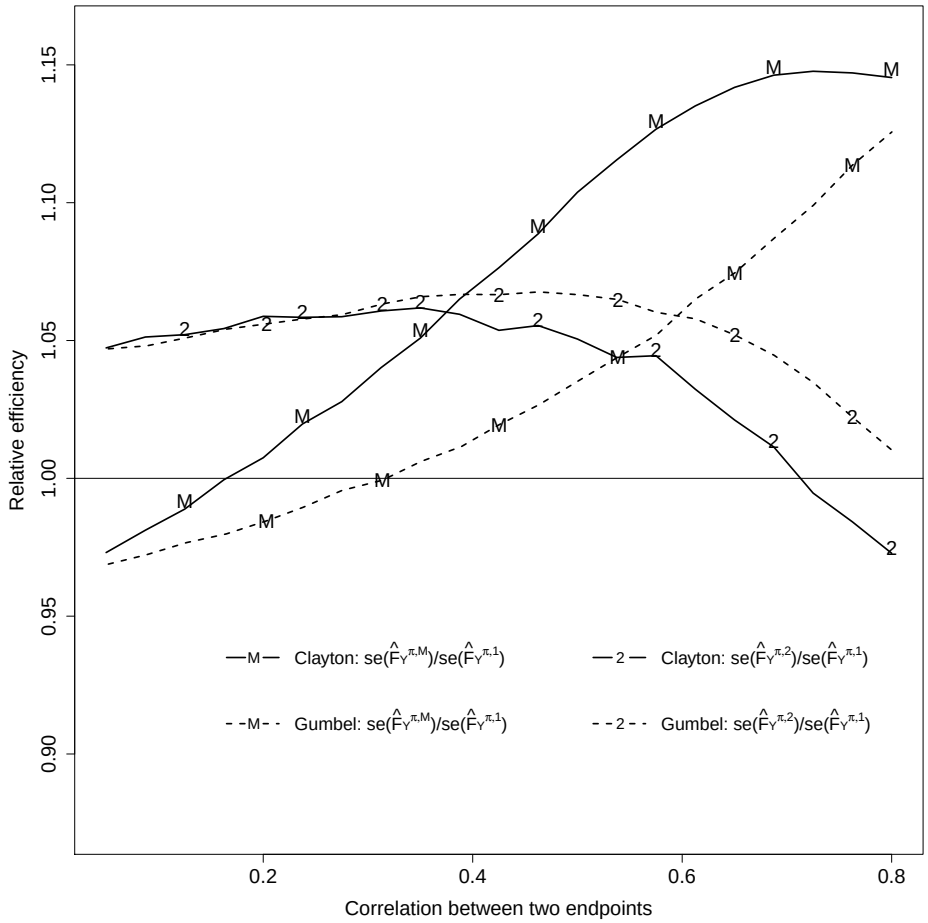

Figure 1: Efficiency comparison for marginal survival estimation

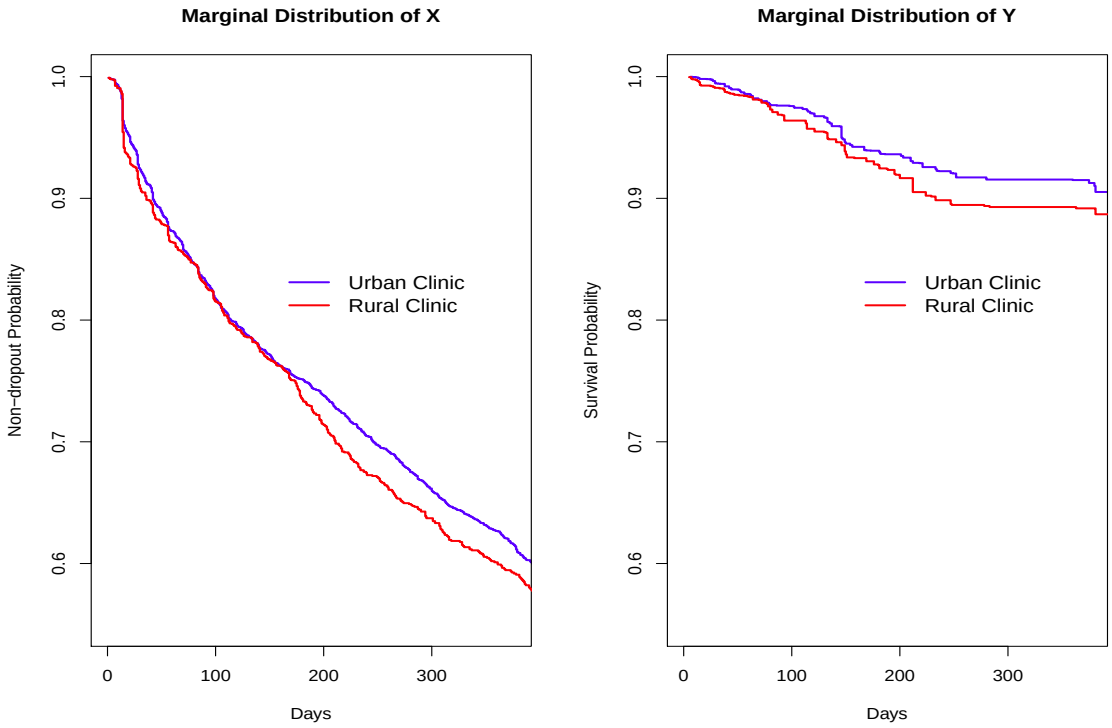

Figure 2: Marginal distribution estimation

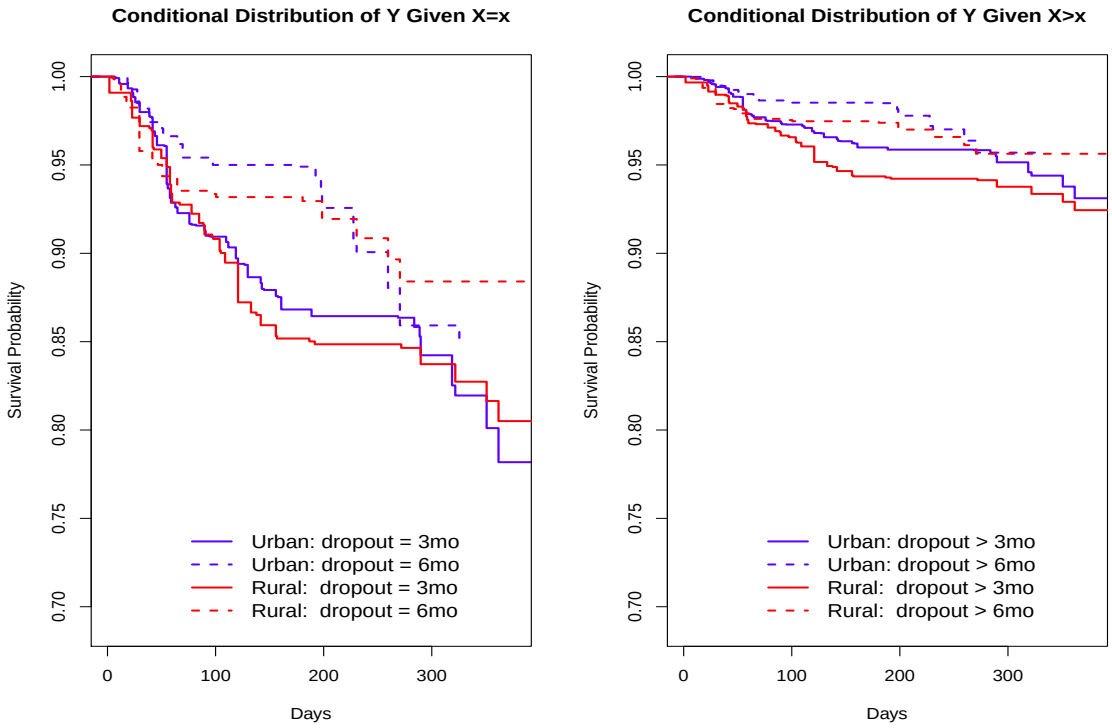

Figure 3: Conditional distribution estimation